Received: 22 May 2023 | Revised: 4 October 2023 | Accepted: 23 November 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12327

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

OSFS-Vague: Online streaming feature selection algorithm based on vague set

Jie Yang[1,2] | Zhijun Wang[2] | Guoyin Wang[2] | Yanmin Liu[1] | Yi He[3] | Di Wu[4]

[1]School of Physics and Electronic Science, Zunyi Normal University, Zunyi, China

[2]Key Laboratory of Big Data Intelligent Computing, Chongqing University of Posts and Telecommunications, Chongqing, China

[3]Department of Computer Science, Old Dominion University, Norfolk, Virginia, USA

[4]College of Computer and Information Science, Southwest University, Chongqing, China

Correspondence

Di Wu.
Email: wudi1986@swu.edu.cn

Funding information

Science and Technology Project of Zunyi, Grant/Award Number: ZSKRPT[2023] 3; Excellent Young Scientific and Technological Talents Foundation of Guizhou Province, Grant/Award Number: QKH-platform talent (2021) 5627; Science and Technology Top Talent Project of Guizhou Education Department, Grant/Award Number: QJJ2022(088); the Guizhou Provincial Department of Education Colleges and Universities Science and Technology Innovation Team, Grant/Award Number: QJJ[2023] 084; National Natural Science Foundation of China, Grant/Award Numbers: 62066049, 62221005, 61936001, 62176070; Department of Education of Guizhou Province, Grant/Award Number: QJJ [2023]084; Science and Technology Foundation of State Grid Corporation of China, Grant/Award Number: 1400-202357341A-1-1-ZN

Abstract

Online streaming feature selection (OSFS), as an online learning manner to handle streaming features, is critical in addressing high-dimensional data. In real big data-related applications, the patterns and distributions of streaming features constantly change over time due to dynamic data generation environments. However, existing OSFS methods rely on presented and fixed hyperparameters, which undoubtedly lead to poor selection performance when encountering dynamic features. To make up for the existing shortcomings, the authors propose a novel OSFS algorithm based on vague set, named OSFS-Vague. Its main idea is to combine uncertainty and three-way decision theories to improve feature selection from the traditional dichotomous method to the trichotomous method. OSFS-Vague also improves the calculation method of correlation between features and labels. Moreover, OSFS-Vague uses the distance correlation coefficient to classify streaming features into relevant features, weakly redundant features, and redundant features. Finally, the relevant features and weakly redundant features are filtered for an optimal feature set. To evaluate the proposed OSFS-Vague, extensive empirical experiments have been conducted on 11 datasets. The results demonstrate that OSFS-Vague outperforms six state-of-the-art OSFS algorithms in terms of selection accuracy and computational efficiency.

KEYWORDS

feature selection, online feature selection, three-way decision, vague set

1 | INTRODUCTION

Feature selection is an efficient method for processing datasets [1–3]. When data volume has increased and data space size is unknown, traditional feature selection methods fail to handle such data well [4]. Online streaming feature selection is developed from the traditional feature selection method [5–9]. The processing of continuous data streaming is the main focus of online streaming feature selection, and it has received a lot of attention recently. Considering data streaming from a real large scale data application, and online streaming feature selection exhibits the capacity to adapt to constantly streaming and changing data. This method is essential for dealing with high-dimensional data since it can process data in real-time.

The ability to flexibly react to various data is the essence of online streaming feature selection. However, most online streaming feature selection algorithms require preset hyperparameters. Different types of datasets necessitate distinct hyperparameters. Appropriate hyperparameters contribute to accelerating the training process, while incorrect parameter settings adversely affect prediction results. For instance, in α-investing [10], setting two hyperparameters in advance before learning is essential to obtain accurate prediction results. However, it is challenging to specify hyperparameters for all datasets. The temporal complexity of some online feature selection methods is too high, which is another problem. Given that the quantity of the datasets used for online streaming feature selection is unknown, the algorithms with high time complexity will consume a lot of time to process large datasets. As a result, the algorithm is only suitable for small datasets. An excellent online feature selection algorithm should both ensure the accuracy of prediction results and control time consumption [11–14].

In recent years, many researchers have conducted experiments related to online streaming feature selection. For example, Wu combined related attributes to remove unnecessary features and improve accuracy while speeding up the learning process [6]. Rough set theory is popular in artificial intelligence and is a useful technique for feature selection and data mining [15–19]. For instance, Eskandari proposed an online streaming feature selection method, named OS-NRRSARA-SA. Based on classical rough set theory, OS-NRRSARA-SA does not require setting any hyperparameters in advance [20]. However, this method fails to directly handle numerical data. To address this limitation, Zhou proposed a new method called OFS-Density by integrating uncertainty theory into online feature selection for both discrete and continuous data [21]. An efficient method for processing feature streaming is OFS-Density based on neighbourhood that does not call for pre-setting hyperparameters. However, OFS-Density has two problems. Firstly, it has a high time complexity, making it difficult to process data with higher dimensions. Secondly, it does not consider the impact of excluding features on subsequent features and fails to reflect the uncertainty in datasets. Causal feature selection has received considerable attention because of its interpretability and predictability, but the current feature selection based on (Markov blanket) MB method cannot deal with streaming features. Therefore, Ou proposes an online streaming feature selection method based on MB [12], named OCFSSFs. This method effectively solves the problems such as long training time and small application range. To speed up the training process, Zhou proposed a new method, named OSFS-ET. After a certain amount of feature training, OSFS-ET speeds up the training process according to active intervening, but it fails to guarantee to search for the optimal results [22].

To improve the application of feature selection methods to data in different environments, the following researchers conducted different studies. Wang presents an Information-theory-based Non-dominated Sorting ACO (called INSA) to improve the multiobjective feature selection to handle the problematic characteristics originated from the feature interactions and highly discontinuous Pareto fronts [23]. Gong proposes a length-adaptive non-dominated sorting genetic algorithm (LA-NSGA) with a length-variable individual encoding and a length-adaptive evolution mechanism for bi-objective high-dimensional feature selection [24]. The new method can introduce to guide individuals to explore promising search space adaptively to solve the problem of too long training time for high-dimensional data. To solve the problem of multimodal MOPs (MMOPs), Han proposed a lot of new methods to improve solution diversity in decision space and performance in objective space [25–27].

To address the issues with the current feature selection method, this paper proposes a unique online streaming feature selection algorithm based on vague set [28], named OSFS-Vague. To combine vague set and three-way decision theories to dynamically analyse the patterns and distributions of streaming features to capture the constantly changing relationships between features and labels [29, 30]. The OSFS-Vague method has the following characteristics.

- OSFS-Vague requires no preset and fixed hyperparameters, making a better selection performance plus the explanations of selected important features.
- Different from the traditional feature selection method, the feature is divided into three parts by the three-way decision theory. In the OSFS-Vague method, features are divided into relevant features, weakly redundant features, and redundant features. The correlation of weakly redundant features is between relevant features and irrelevant features, so we reserve weakly redundant features. Finally, weakly redundant features compare the features in the optimal feature subset, and the inferior features are replaced. This classification improves prediction accuracy and is more consistent with the cognitive process of the real world for humans.
- A novel method is designed to obtain the relationship between features and labels based on vague set. Different from labels calculated separately, OSFS-Vague combines positive and negative evidence to describe the features. The correlation and redundancy between features are described from multiple perspectives to improve training speed while ensuring prediction accuracy.

To sum up, this paper starts with uncertainty theory and integrates Vague and three-way decision theory into the feature selection framework. The feature screening method is improved by vague set theory. Then, the three-way decision theory is used to improve the process of obtaining the optimal feature subset, so that the decision process is more consistent with the human cognitive process. This new method can effectively improve the accuracy and training speed of feature screening, and expand the application of uncertainty theory in the selection of stream features. Subsequent experiments will prove the effectiveness of this method.

The remainder of the paper is organised as follows: Section 2 discusses related work. Section 3 presents a brief introduction to uncertain theory and a new method based on vague set for online streaming feature selection. Section 4 reports experimental results. Lastly, Section 5 concludes the paper.

2 | RELATED WORK

In this section, we review various representative traditional feature selection methods and online streaming feature selection methods.

2.1 | Traditional feature selection methods

The key step of data preprocessing is feature selection that offers a wide range of benefits, such as reducing the dimensionality of the datasets, improving model training speed, preventing over-fitting, and minimising processor requirements in data processing and analysis [31].

There are several feature selection types, such as filter, wrapper, and embedded [32–34]. The filter method is a technique that assesses features in isolation from the model and classifier, and takes into account their correlation with other features. This approach allows for efficient filtering and decreases the computational burden. However, it may sacrifice some accuracy. In contrast, the wrapper method tightly integrates feature evaluation with the learning algorithm, resulting in higher accuracy, but with increased computational cost. In the embedded, the model learning training process is connected with the feature selection procedure. It is reducing training costs, but the results may be more biased towards the classifier used for training.

The most traditional feature selection only focuses on the relationship between individual features and labels. More specifically, ReliefF calculates the weight of each feature by finding its neighbours from different samples [35]. Laplacian Score determines the weight of features according to the fluctuation of the Euclidean distance sample value [36]. Fisher Score calculates the weight of each feature by the ratio of inter-class separation and intra-class differentiation [37]. Mutual Information (MI) measures the independence of data that was introduced. It primarily takes into account the distribution of a particular feature to other features [38]. Information Networks Feature (INF) is an unsupervised filtering method proposed and it treats each feature as a node, and multiple nodes are combined to form a graph [39]. The more node is connected to other nodes, the higher the corresponding node score, indicating the more important the feature is. Tsai analysed the effect of combining multiple feature selection algorithms [40]. This study categorised the types of combinations into three groups: nine parallel combinations, and nine serial combinations.

To improve the application range and prediction accuracy of the feature selection method, researchers have applied new search methods based on the traditional methods. The whale optimization algorithm (WOA) has low population diversity and a poor search strategy. M.H. Nadimi-Shahraki adopts a pooling mechanism and three effective search strategies to overcome these problems, named BE-WOA [41]. The current diagnostic methods are too single to effectively diagnose the coronavirus disease 2019 (COVID-19). Thus, Hu constructed a new framework by exploring problems such as the slight appearance difference between mild cases and severe cases, the interpretability, the High Dimensional and Low Sample Size (HDLSS) data, and the class imbalances, named MM-SVM [42]. Incremental feature selection can retain the previous training results to update the optimal feature set based on the added-in data. However, this method has too many redundant calculations, which reduces the training speed and wastes memory. Therefore, Yang proposed a new method to solve repetitive computation based on sample selection and a feature-based accelerator, named IFS-SSFA [43]. Zhou proposed a new balanced spectral feature selection (BSFS) method based on the traditional unsupervised spectral feature selection method. This method can obtain optimal features and also reveal the balanced structure of data [44].

Traditional feature selection has a good performance on traditional datasets, but it is not suitable for datasets with unknown sizes. Because traditional feature selection will spend more time waiting for the arrival of the feature. Moreover, when the dataset is updated, traditional feature selection needs to be retrained while online streaming feature selection continues training according to the last results to avoid wasting time.

2.2 | Online feature selection methods

Online streaming feature selection can effectively to cope with high-dimensional data, so it has attracted many scholars' attention in recent years [45, 46].

Online feature selection algorithms not only calculate the relationship between features and labels but also focus on the relationship between features. Therefore, the training time for online streaming feature selection is longer than that of traditional feature selection, but the prediction accuracy is higher because redundant features are excluded.

In recent years, there have been efforts to tackle online streaming feature selection. For instance, Zhou et al. proposed α-investing, which requires pre-set hyperparameters [10]. The effectiveness of the algorithm and the number of feature selection rely heavily on the chosen parameter settings.

Although the training speed of α-investing is very fast, the prediction accuracy is low. Therefore, Wu et al. introduced two online streaming feature selection algorithms called OSFS and Fast-OSFS [6], which effectively improved the prediction accuracy. The algorithm has two key steps.

1) Relevance analysis, which is to exclude irrelevant features.
2) Redundancy analysis, which eliminates redundant features.

OSFS relies on conditional uncertainty to select features, which results in the issue of needing a lot of training instances. OSFS generates unreliable results when the size of the datasets is limited.

OSFS and Fast-OSFS have better prediction accuracy than other online stream feature selection methods, but the training time is very long and the efficiency of processing data sets is low. Yu et al. presented SAOLA, a highly scalable feature selection method [4]. SAOLA has advantages in data processing speed and is often used to process high-dimensional datasets. SAOLA sets a threshold to determine the relevance between two features.

SAOLA can't deal with redundant features effectively, so the number of features obtained after feature selection is too large and the prediction accuracy is low. To solve this problem, Zhou et al. introduced an OFS-Density, an online streaming feature selection method based on neighbourhood rough set [21]. The algorithm calculates the correlation of the new arriving feature through the neighbourhood relation. When a new feature arrives, it will be chosen if its correlation is higher than the average correlation. When a feature is selected, the algorithm will judge and remove redundant features. The algorithm incorporates uncertainty theory to effectively eliminate redundant features and train faster.

OSFS-Density proves the effect of rough set in streaming feature selection, but there is no systematic analysis of the relationship between rough set and streaming feature selection framework. Thus, Zhou et al. propose a generalised assembly rough set-based framework for streaming feature selection, named RS-SFSF [15]. This method can measure the selected features as integral without any domain knowledge.

To make feature selection consistent with real-world cognitive processes, we put forward OSFS-Vague by integrating three-way decision theory into online streaming feature selection. OSFS-Vague overcomes the shortcomings of current streaming feature selection methods without grasping the global features. Meanwhile, the vague set theory is used to describe the relationship between features and labels from positive and negative aspects, evaluating the accuracy of feature description in a more comprehensive way. Therefore, the OSFS-Vague method can effectively improve the prediction accuracy.

In the next section, we will improve the original framework model of feature selection and integrate vague set and three-way decision into correlation and redundancy analysis.

3 | ONLINE STREAMING FEATURE SELECTION METHODS BASED ON VAGUE SET

First, we cover some fundamental definitions in this section. Then describe the algorithm in three parts and illustrate the overall algorithm training process through an example.

3.1 | Symbols and notations

The adopted symbols of this article are summarised and explained in Table 1.

TABLE 1 SYMBOL annotations.

Symbol	Annotations
U	The set of objects to be discussed within a range is called the domain.
B	B is relation of equivalence on U.
V	A vague set on U.
$t_V(x)$	A truth-membership function on V.
$n_V(x)$	A false-membership function on V.
A	A fuzzy set on U.
λ	Control the stringency of filtering feature on online streaming feature selection.
D	A data set on U.
M	All samples in the feature set.
F	A feature set on U.
f	An item of F.
$f_{M,n}$	A vector corresponding to M instances.
$\mu_A(x)$	Represents the degree to which element x conforms to the definition of set A.
O	Attribute values for sets K and L.
I	The feature set at time t.
W	A weakly redundant feature set on F.
K	The set of all the different feature sets.
L	A label set on F.
l	An item of L.
G	The set is sorted by L.
g	An item of G.
S	An optimum feature set on F.
T_x	The quantity of object sets with x as a label value that is consecutive.
num_x	The quantity of labels having x as their value.
∂	A function that provides information about each object's attribute value.
T	A feature streaming on U.

3.2 | Basic related definitions

According to their feature values for attribute B, objects are grouped into equivalence classes in the traditional rough set model [47], denoted by $[x]_B$. $\{[x_i]_B | x_i \in U\}$ denotes a system to describe an arbitrary subset of the sample space, where $U = \{x_1, x_2, \ldots, x_n\}$ is a non-empty finite set of objects called the universe. Rough set theory is often applied to data mining and classification tasks. For subset X, the following are the definitions of lower and upper approximation.

$$\underline{B}X = \{[x_i]_B | [x_i]_B \subseteq X, x_i \in U\}, \tag{1}$$

$$\overline{B}X = \{[x_i]_B | [x_i]_B \cap X \neq \varnothing, x_i \in U\}. \tag{2}$$

The maximum union of granules in X is represented by the lower approximation, while the minimum union of granules in

X is represented by the upper approximation. Then, the objects in U can be divided into three parts, called the positive regions, boundary regions and negative regions.

$$POS_B(X) = \underline{B}X, \tag{3}$$

$$BND_B(X) = \overline{B}X - \underline{B}X, \tag{4}$$

$$NEG_B(X) = U - \overline{B}X. \tag{5}$$

The vague set has a more capacity to handle uncertain information than the fuzzy set as a traditional soft computing tool. Vague value is characterised by a truth-membership function $t_V(x)$ and a false-membership function $n_V(x)$. The evidence for x delivers the lower bound $t_V(x)$ of the grade of membership of x, while the evidence against x delivers the lower bound $n_V(x)$ of the negation of x. The vague set model supports handling both continuous and discrete datasets. This section briefly discusses relative basic concepts.

Definition 1 (Fuzzy set) [48]: Let U denote a universe of discourse. Then a fuzzy set A in U is defined as a set of ordered pairs $A = \{< x, \mu_A(x) > | x \in U\}$, where μ_A: $U \rightarrow [0, 1]$ is the membership function of A and μ_A is the grade of belongingness of x in A.

Definition 2 (Vague set) [28]: A vague set V in U is characterised by a truth-membership function $t_V(x)$ and a false-membership function $n_V(x)$, $t_V(x)$ is a lower boundary on the grade of membership of x derived from the evidence for x, and $n_V(x)$ is a lower boundary on the negation of x derived from the evidence against x. Both $t_V(x)$ and $n_V(x)$ are associated with a real number in the interval [0,1] with each point in U, where $t_V(x) + n_V(x) \leq 1$. That is, $t_V(x)$: $U \rightarrow [0, 1]$ and $n_V(x)$: $U \rightarrow [0, 1]$.

If U is continuous, a vague set V is depicted as follows:

$$V = \int_U [t_V(x), 1 - n_V(x)]/xdx. \tag{6}$$

If U is discrete, a vague set V is depicted as follows:

$$V = \sum_{i=1}^{n} [t_V(x_V), 1 - n_V(x_V)]/xi. \tag{7}$$

Here, $t_V(x) \leq 1 - n_V(x)$, $1 \leq i \leq n$. When $t_V(x) = 1 - n_V(x)$, the vague set will transform to the fuzzy set. Fuzzy set is a special vague set.

Definition 3 (Streaming Features) [6]: Assume a dataset D, which has M instances and a feature set K. $K = \{F_1, F_2, \ldots, F_N\}$ where $F_n = \left[f_{1,n}, f_{2,n}, \ldots, f_{M,n}\right]^T$, $n \in \{1, 2, \ldots, N\}$ is a vector corresponding to M instances. The features are obtained gradually with time.

Definition 4 (Online streaming feature selection) [6]: $I_{n-1} = \{F_1, F_2, \ldots, F_{n-1}\}$ be obtained from streaming features set at time n-1. We also obtain the optimal feature set S_{n-1} from I_{n-1} by selection feature at time n - 1, where $S_{n-1} \subseteq I_{n-1}$.

Definition 5 (Distance Correlation coefficient) [49]: Suppose two objects X and Y, where $X = [x_1, x_2, \ldots, x_n]$, $Y = [y_1, y_2, \ldots, y_n]$. *dcorr* (X, Y) is the value of distance correlation of X and Y. *dcorr* (X, Y) is defined as follows:

$$dcorr(X, Y) = \frac{dcov(X, Y)}{\sqrt{dcov(X, X)dcov(Y, Y)}}. \tag{8}$$

Where, $dcov^2(X, Y) = S^{\wedge}{}_1 + S^{\wedge}{}_2 - 2S^{\wedge}{}_3$. $S^{\wedge}{}_1, S^{\wedge}{}_2$ and $S^{\wedge}{}_3$ is defined as follows.

$$\underset{1}{S^{\wedge}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| x_i - x_j \right\|_{dX} \left\| y_i - y_j \right\|_{\mathrm{dY}}, \tag{9}$$

$$\underset{2}{S^{\wedge}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| x_i - x_j \right\|_{\mathrm{dX}} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| y_i - y_j \right\|_{\mathrm{dY}}, \tag{10}$$

$$\underset{3}{S^{\wedge}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \left\| x_i - x_l \right\|_{\mathrm{dX}} \left\| y_j - y_l \right\|_{\mathrm{dY}}. \tag{11}$$

3.3 | A new definition of correlation

Definition 6 (Regular value of feature): Given an object of finite and non-empty U. $F = \{f_1, f_2, \ldots, f_m\}$ is a feature set, and L is the label set. Different features correspond to different Regular values of features. Sorting the samples according to the value of the current feature. $L = \{l_1, l_2, \ldots, l_n\}$ is the label value after sorting. $G = \{g_1, g_2, \ldots, g_n\}$ is the regular value for l_i. G is derived from both positive and negative evidence functions. When we have chosen one label as positive evidence, other labels are negative evidence. Evidence function is defined as follows:

$$t_V = 1 - \frac{T_x}{num_x}, \tag{12}$$

$$1 - n_V = 1 - \frac{T_y}{num_y}. \tag{13}$$

T_x represents the number of sets of objects with a consecutive label value of x and num_x represents the number of labels with a value of x. If x is 0, T_x more approaches 0, which indicates that the current feature is more useful for classifying. Conversely, if T_x more approaches num_x, it indicates that the current feature is more unuseful for classifying. Finally, we integrate the regular value of the current feature, where $g_i = [t_V(l_i), 1 - n_V(l_i)]$. Table 2. Shows an example.

Features f_1 to f_4 have 10 samples (x_1 to x_{10}) and a label value (0 or 1).

Firstly, we sort the sample value, $\triangle f_1$= {0.5,1.1,1.2,2.5,3.6, 4.2,5.1,8.4,9.8,12.3} and o_1 = {0,0,0,1,1,0,1,1,0,0}. We assume that x is 0 in Equation (12), and y is one in Equation (13). Thus, $T_x = 3$, $num_x = 6$. Similarly, $T_y = 2$, $num_y = 4$. Finally, the regular value is $g_1 = [0.5, 0.5] = 0.5$. Based on this, the regular values of all features are calculated as in Table 3.

It is obviously that $g_4 > g_2 > g_1 > g_3$.

3.4 | A new algorithm

In this study, we suggest a unique strategy for online streaming feature selection that is based on vague set [28], called OSFS-Vague. This method has the following characteristics.

TABLE 2 AN example dataset.

$x \in U$	f_1	f_2	f_3	f_4	L
x_1	1.2	−1.2	−1.3	−9.9	0
x_2	5.1	7.6	−1.3	−6.4	0
x_3	12.3	−2.1	12.5	1.6	0
x_4	0.5	5.7	7.5	−8.8	0
x_5	8.4	7.8	9.6	−4.2	1
x_6	3.6	4.3	3.6	−1.5	1
x_7	2.5	4	−1.7	−5.1	1
x_8	1.1	−5.5	2.7	3.2	0
x_9	4.2	−2.4	−5.6	1.8	0
x_{10}	9.8	2	−1.5	−7.9	0

TABLE 3 REGULAR value of different features.

	f_1	f_2	f_3	f_4
T_x	3	1	4	1
num_x	6	6	6	6
T_y	2	1	3	0
num_y	4	4	4	4
g_i	0.5	0.82	0.27	0.99

1) Existing methods tend to focus only on the relationship between features and labels, ignoring the relationship of different label values. OSFS-Vague calculates the correlation of different features with label value and uses vague set theories to decide the advantage of a feature from multiple perspectives rather than from a single perspective.
2) Existing some methods use the Pearson Correlation Coefficient (PCC) to describe the relevance between different features [50]. However, PCC is only effective for linearly describe dependent data. OSFS-Vague combines the vague set with the distance correlation coefficient [49], which solves the problem of linearly independent data, thus improving the accuracy of excluding redundant features.
3) OSFS-Vague compares the weakly redundant feature set with the optimal feature set to exclude poor features. Based on vague set theories [28, 51], OSFS-Vague works well with real-world datasets and doesn't require any hyperparameters to be supplied before training.

In this paper, we extend the traditional framework by weak redundancy analysis and the three-way decision theories [29, 30]. The specific procedure is shown in Figure 1. First, when a new feature arrives, we calculate its correlation by Defining 6, and Equation (14) to classify features into relevant and irrelevant features [6, 52–54]. Second, relevant features need to be analyzed by Definition 7 and Definition 8, and the feature will be divided into optimal features, weakly redundant features, and redundant features. Finally, we compare the optimal feature set with the weakly redundant features, and the poor features in the optimal feature set are replaced by the weakly redundant features. Each step will be discussed in detail as follows.

3.4.1 | Correlation analysis

To select the highly correlation features from the feature streaming, we calculate g for each feature and compare with $\Re$. If gis greater than $\Re$, we consider the feature is highly correlation.

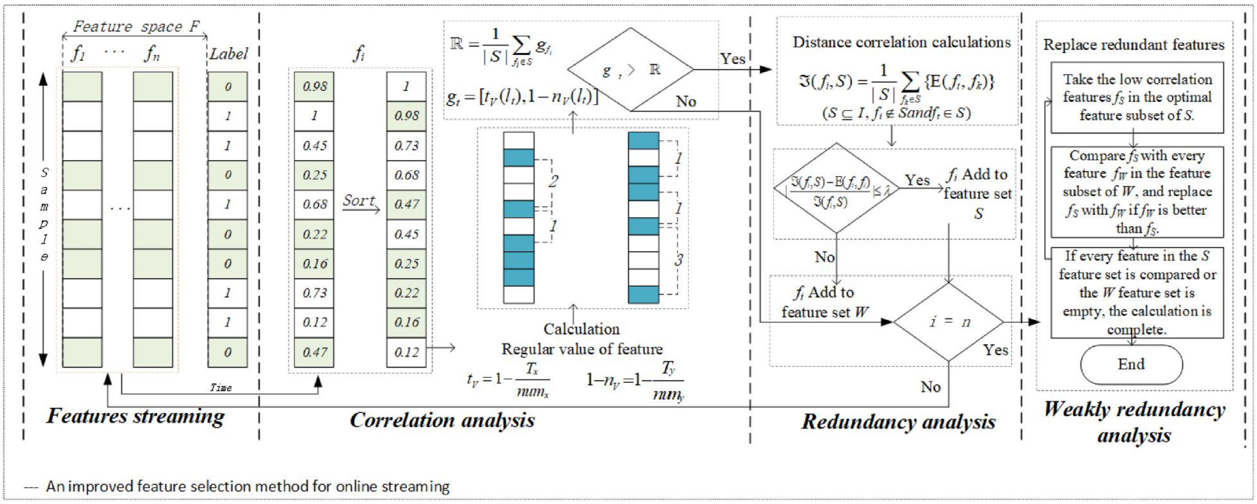
Feature space F
f_1 · · · f_n
Label
Sample
Time
Features streaming
f_i
0.98
1
0.45
0.25
0.68
0.22
0.16
0.73
0.12
0.47
Sort
1
0.98
0.73
0.68
0.47
0.45
0.25
0.22
0.16
0.12
$\mathbb{R} = \frac{1}{|S|}\sum_{f_i \in S} g_{f_i}$
$g_t = [t_V(l_t), 1 - n_V(l_t)]$
$g_t > \mathbb{R}$
Yes
No
2
1
1
1
3
Calculation
Regular value of feature
$t_V = 1 - \frac{T_x}{num_x}$
$1 - n_V = 1 - \frac{T_y}{num_y}$
Correlation analysis
Distance correlation calculations
$\Im(f_i, S) = \frac{1}{|S|}\sum_{f_k \in S}\{E(f_i, f_k)\}$
$(S \subseteq I, f_i \notin S and f_t \in S)$
$\left|\frac{\Im(f_i,S) - E(f_i,f_t)}{\Im(f_i,S)}\right| \le \lambda$
Yes
f_i Add to feature set S
No
f_i Add to feature set W
$i = n$
Yes
No
Redundancy analysis
Replace redundant features
Take the low correlation features f_S in the optimal feature subset of S.
Compare f_S with every feature f_W in the feature subset of W, and replace f_S with f_W if f_W is better than f_S.
If every feature in the S feature set is compared or the W feature set is empty, the calculation is complete.
End
Weakly redundancy analysis
--- An improved feature selection method for online streaming

FIGURE 1 A new framework for online streaming feature selection.

There is a feature streaming as $F = \{f_1, f_2, \ldots, f_n\}$ and the optimal feature set S, $S \subseteq F$. At time t, the arriving feature is f_t. If $g_t > \Re$, f_t is high correlation feature. Where, $\Re$ is defined as follow:

$$\Re = \frac{1}{|S|}\sum_{f_i \in S} g_t. \tag{14}$$

Algorithm 1 illustrates the precise filtering method of feature selection.

Algorithm 1 Correlation calculation

```
Input: $X$ (Sample value), Y (Label value),
       C (Total number of classes)
Output: Average correlation of feature
Let $Total\_a = 0$, $Total\_b = 0$, $result = 0$.
for each $i$ in $C$
    $Total\_a$ = calculate the correlation
      of label $i$ by $Y$
    for $c$ range ($i + 1$, $C + 1$)
      $temper$ = calculate the correlation
        of label $c$ by $Y$
      $Total\_b = Total\_b + temper$
    end for
    $Total\_b = Total\_b / (C - i)$
    $result = result + int\ (Total\_a, Total\_b)$
end for
return $result / (C - 1)$
```

3.4.2 | Redundancy analysis

Redundancy analysis is the calculation of the degree of similarity between different features. If two features describe concepts that are very similar and one of them needs to be removed. If the feature that needs to be removed belongs to the optimal feature set, delete it directly. Otherwise, we remove it and add it to the weakly redundant feature set.

To measure the redundant relationship between each feature, the current feature is contrasted with the features in the optimal feature set.

Definition 7 (Distance correlation analysis): Assume a feature set $F = \{f_1, f_2, f_3, \ldots, f_n\}$, and the current feature set is $I = \{f_1, f_2, f_3, \ldots, f_t\}$, $F \supseteq I$. The current feature is f_t. $dcorr(f_t, f_i)$ to calculate the distance correlation between f_t and f_i, $f_i \in I$ ($i = 1, 2, 3\ldots, t-1$).

$$\Im(f_i, S) = \frac{1}{|S|}\sum_{f_k \in S}\{dcorr(f_t, f_k)\}(S \subseteq I, f_i \notin S \text{ and } f_t \in S).$$

If the distance correlation between f_t and all other features is smaller than the distance correlation between f_i and other features, then the current feature is not redundant.

In the real world, many results are not so strict. As long as the result fluctuates within range λ, we think the result is correct. λ is set to 0.05 as the default value.

$$\left|\frac{\Im(f_i, S) - dcorr(f_t, f_i)}{\Im(f_i, S)}\right| \leq \lambda \tag{15}$$

Definition 8 (Regional subdivided of the three-way decision on feature streaming): Given a decision system $p = (U, K \cup L, O, \partial)$. $F_1 \subseteq F_2 \subseteq \ldots \subseteq F_n \subseteq K$ and T is a feature streaming on U. O is the range of F and L, ∂ is an information function that specifies the value of each object in U. Following is a denotation for the three disjoint regions:

$$POS_{F_i}(T) = \left\{f \in U | g_i > \Re, \left|\frac{\Im(f,S) - dcorr(f_i, f)}{\Im(f,S)}\right| \leq \lambda\right\},$$
$$BND_{F_i}(T) = \{f \in U | g_i > \Re\},$$
$$ENG_{F_i}(T) = \{f \in U | g_i < \Re\}.$$

Combining Equation (14) and (15). If f_i is partitioned into $POS_F(T)$, the feature considered is highly correlation. If f_i is partitioned into $ENG_F(T)$, the feature is considered to be the irrelevant feature. If f is partitioned into $BND_F(T)$, the feature is considered to be the weakly redundant feature, and the judgment is postponed. The change of regions is shown in Figure 2.

The specific filtering method is shown in Algorithm 2.

3.4.3 | Weakly redundant analysis

In the three-way decision, the features in the boundary domain may be divided into positive or negative domains with the change of conditions. Therefore, it is necessary to filter again at the end of streaming feature selection.

The filtered feature set falls between the optimal feature set and the weakly redundant feature set, which has filtered most of the irrelevant features, but some features can still be optimised. The three-way decision theories appear in the delayed decision process [29, 30, 54]. The subsequent specific filter method is shown in Algorithm 3.

Algorithm 2 Distance correlation calculation

```
Input: $F\_1$ and $F\_2$ is condition features,
       $C$ (Total number of classes)
Output: Distance correlation between
   features
Let $Total\_a = 0$, $Total\_b = 0$, $result = 0$.
for each $i$ in $C$
    $Total\_a$ = calculate the distance
      correlation between
F_1 and F_2 on label $i$
    for $c$ in range ($i + 1$, $C + 1$)
      $temper$ = calculate the distance
        correlation
between F_1 and F_2 on label $c$
      $Total\_b = Total\_b + temper$
    end for
    $Total\_b = Total\_b / (C - i)$
    $result = result + int\ (Total\_a, Total\_b)$
end for
return $result / (C - 1)$
```

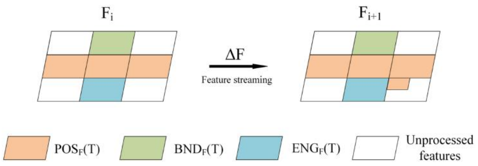

FIGURE 2 The feature set is subdivided into feature regions of OSFS-Vague.

Algorithm example and analysis. A more specific example illustrates the framework proposed in this paper. We assume a feature set $F = \{f_1, f_2, f_3, \ldots, f_n\}$ and the optimal feature set S, $S \subseteq F$. At time t, we will obtain a feature f_t. Firstly, the correlation of the feature in S with f_t is compared by Algorithm 1. If the *Mean* is greater than the correlation of f_t, we think f_t is an irrelevant feature. Secondly, the distance correlation between f_t and each feature in S is calculated by Algorithm 2. If the distance correlation between any features in S and f_t is smaller than the distance correlation between S, f_t is considered a non-redundant feature, otherwise, f_t is considered a weakly redundant feature. Finally, when the feature selection ends. Reevaluating the feature correlation in the optimal feature set. Because with the increase of features, the feature may not meet the requirements of the optimal feature set. Features that are less than the average correlation of the optimal feature are compared with the weakly redundant features and replaced with the current feature if there are better features in the weakly redundant features than the current feature.

OSFS-Vague can calculate the correlation between features and labels to obtain features for classifying data sets. At the same time, the correlation between different features is obtained to eliminate redundant features and further improve the prediction accuracy. OSFS-Vague algorithm can get the best features beneficial to data set classification, and the fast training speed is suitable for processing large data sets. In the experiment in Section 4, the ability of OSFS-Vague to process classified data sets is verified.

Time complexity analysis. The time complexity of OSFS-Vague is divided into three parts. Firstly, we suppose a dataset has m samples, n features, and c labels. When new features arrive with time, we sort them and do linear processing, the complexity is O (n × (m^2 + a)). Secondly, the features with high correlation needs to require redundancy analysis. Time spent in feature computation redundancy is related to the quantity of features in the optimal feature set. The complexity is O (1/2 × n × b) in the worst-case scenario. Finally, in the worst case, we recalculate all the features for weakly redundant analysis, and the time complexity is O (1/4 × n^2 × b). The total time complexity is O (m^2 + n^2).

Algorithm 3 Filter weakly redundant feature set

```
Input: $S$ (Optimum feature set), $W$ (Weakly redundant feature set), $mean$ (Average correlation of optimum feature set)
Output: Optimum feature set
Let $\lambda = 0.05$, $f_S = \varnothing$, $f_W = \varnothing$ $S\_dis = 0$, $W\_dis = 0$
for each $f_S$ in $S$
    if ($W = \varnothing$)
        break
    end if
    $Cof_S$ = calculate the correlation of $f_S$
    if ($Cof_S < mean$)
        continue
    end if
    for each $f_W$ in $W$
        $Cof_W$ = calculate the correlation of $f_W$
        if ($Cof_W < mean$)
            continue
        end if
        $S\_dis$ = calculate the distance correlation by $f_S$
        $W\_dis$ = calculate the distance correlation by $f_W$
        if ($W\_dis < S\_dis$ or abs($W\_dis$ - $S\_dis$) / $S\_dis$ <= $\lambda$)
            $S$.remove($f_S$)
            $S$.add($f_W$)
            $W$.remove($f_W$)
        end if
    end for
end for
return S
```

4 | EXPERIMENTS

In the experiments, we mainly solve the following research questions (RQs):

RQ 1. Does OSFS-Vague outperform other algorithms in the following test settings?

RQ 2. What are the influences of hyper-parameter λ and weakly redundant features for OSFS-Vague?

4.1 | General settings

Datasets. In this paper, we have selected 10 benchmark datasets from various sources, including DNA microarray datasets [55] and UCI repositories [56]. Datasets have different types and varying sizes. Datasets with large samples are mainly used to test the efficiency of the algorithm in processing features. Large feature sets are typically used to test the number of feature selections and the precision of predictions. Table 4 lists the different datasets used in the experiment.

Baselines. The algorithms are compared mainly by the prediction accuracy. Prediction accuracy refers to the classification tendency of the classifier after being trained by different algorithms and then compared with the actual results of the test prediction classification results. The following algorithms

TABLE 4 Details of selected datasets.

Dataset	Instance	Feature	Class
TOX_171	171	5748	4
Lymphoma	62	4026	3
Leukemia	72	7129	2
Parkinson disease classification	756	753	2
Colon	62	2000	2
Lung cancer	181	12,533	2
Gas sensor array under flow modulation	53	432	4
Prostate	102	6033	2
Lung	203	3312	5
DLBCL	77	5469	2
Arcene	200	10,000	2

are used in this paper: ReliefF [35], Fisher [37], MI [38], PCC (Pearson Correlation Coefficient) [57], Laplacian [36], OFS-Density [21], OSFS [6], Fast-OSFS [6], SAOLA [4], α-investing [10] and NRS-SFSF [15]. Table 5 displays the specifics of the algorithms.

Implementation Details. We conduct statistical tests using the Friedman test (F-rank) and the Wilcoxon Signed-Ranks test based on prediction accuracy results [58]. The algorithm performs better the lower the Rank value. If the p-value is less than 0.05, it indicates that there are significant differences among the tested algorithms. We use K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Gradient Boosting Regression Tree (GBRT) results as results of prediction accuracy. The experiments are performed on a computer with 16 GB RAM, 2.3 GHz, and Intel (R) i7-11,800H processor.

4.2 | OSFS-Vague and other algorithm comparison test (RQ. 1)

In this part, the feature selection number, prediction accuracy, and consumption time of OSFS-Vague are compared to those of other approaches.

4.2.1 | Prediction accuracy for traditional feature selection

To verify that OSFS-Vague is more advantageous than other traditional feature selection algorithms, we conducted precision prediction tests between OSFS-Vague and other feature selection algorithms under three different classifiers. The prediction accuracy test refers to the optimal feature set to predict labels of different samples in the classifier. The higher the accuracy, the better the algorithm. Hyperparameters of KNN, SVM, and GBRT classifiers are all the same for different algorithms in the same datasets.

TABLE 5 Descriptions of all the involved models.

Model	Description
ReliefF [28]	The correlation between feature and label is based on the feature's ability to distinguish near samples.
Fisher [30]	The algorithm obtains the correlation of features by the ratio of inter-class separation and intra-class differentiation.
MI [31]	This method is based on information entropy and only supports the calculation of discrete variables.
PCC [50]	This method is used to detect the degree of linear correlation between two continuous variables. The greater the difference in values, the higher the degree of linear correlation.
Laplacian [29]	The method determines the weight of features according to the fluctuation of the Euclidean distance sample value.
OFS-Density [19]	It takes neighbourhood relations into account to describe the relationship between the same feature and different features.
OSFS [6]	A traditional feature selection method that compares the scores of different features to get the best selection.
Fast-OSFS [6]	This method is an advanced version of OSFS, which optimises the selection features and can speed up most of the data training process.
SAOLA [4]	Compared with other algorithms, this algorithm has the advantage of training speed, but the prediction accuracy is not high.
α-investing [10]	Compared with other algorithms, this algorithm requires to pre-set additional hyperparameters.
NRS-SFSF [13]	Compared with other algorithms, this algorithm requires to pre-set additional hyperparameters.
OSFS-Vague	A novel algorithm based on vague set theories is proposed in this paper.

Table 6 lists the prediction accuracy for different algorithms in 11 datasets. In particular, the win-loss-tie ratio of OSFS-Vague is 23/3/7, which is the greatest advantage compared with Fisher. On the contrary, compared with ReliefF, OSFS-Vague has little advantage, which the win-loss-tie ratio is 19/5/9. OSFS-Vague is superior to other algorithms in average prediction accuracy and F-rank value. Although the all predction accuracy of OSFS-Vague is not optimal, OSFS-Vague is superior to other algorithms in general.

4.2.2 | Feature number analysis for online feature selection

The number of features will affect the time of data processing by the classifier, thus affecting the overall efficiency.

From Table 7. The average number of features obtained by OSFS-Vague, OFS-Density, OSFS, Fast-OSFS, and NRS-SFSF are similar, and the results from various datasets are not significantly different. The number of features obtained by

TABLE 6 Prediction accuracy of OSFS-Vague vs traditional feature selection algorithms.

Dataset	Type	OSFS-Vague	ReliefF	Fisher	MI	PCC	Laplacian
Lung cancer	KNN	0.827	0.827	0.827	0.827	0.827	0.827
	SVM	0.172	0.827◦	0.827◦	0.827◦	0.827◦	0.827◦
	GBRT	0.979	0.861•	0.769•	0.726•	0.902•	0.756•
TOX_171	KNN	0.825	0.610•	0.573•	0.510•	0.667•	0.554•
	SVM	0.641	0.477•	0.264•	0.264•	0.422•	0.264•
	GBRT	0.489	0.322•	0.250•	0.220•	0.323•	0.244•
Lymphoma	KNN	0.672	0.672	0.549•	0.672	0.672	0.672
	SVM	0.672	0.672	0.672	0.672	0.672	0.672
	GBRT	0.979	0.721•	0.615•	0.491•	0.873•	0.551•
Gas sensor array under flow modulation	KNN	0.926	0.498•	0.823•	0.386•	0.528•	0.498•
	SVM	0.776	0.344•	0.667•	0.344•	0.344•	0.344•
	GBRT	0.762	0.448•	0.663•	0.220•	0.551•	0.379•
Prostate	KNN	0.900	0.514•	0.485•	0.549•	0.485•	0.537•
	SVM	0.514	0.514	0.485•	0.514	0.485•	0.485•
	GBRT	0.864	0.846•	0.494•	0.514•	0.767•	0.554•
Lung	KNN	0.683	0.683	0.683	0.683	0.029•	0.683
	SVM	0.683	0.683	0.683	0.683	0.683	0.683
	GBRT	0.696	0.631•	0.557•	0.423•	0.557•	0.410•
DLBCL	KNN	0.250	0.749◦	0.749◦	0.749◦	0.250	0.749◦
	SVM	0.749	0.749	0.750◦	0.749	0.250•	0.749
	GBRT	0.907	0.733•	0.726•	0.615•	0.776•	0.693•
Leukemia	KNN	0.788	0.943◦	0.647•	0.647•	0.647•	0.647•
	SVM	0.760	0.943◦	0.746◦	0.647•	0.915◦	0.647•
	GBRT	0.743	0.905◦	0.703•	0.655•	0.916◦	0.651•
Parkinson disease classification	KNN	0.923	0.479•	0.932•	0.850•	0.935◦	0.949◦
	SVM	0.746	0.746	0.746	0.746	0.746	0.746
	GBRT	0.750	0.746•	0.749•	0.746•	0.746•	0.746•
Colon	KNN	1	0.926•	0.930•	0.885•	0.955•	0.910•
	SVM	0.868	0.639•	0.717•	0.639•	0.852•	0.639•
	GBRT	0.822	0.565•	0.676•	0.569•	0.778•	0.560•
Arcene	KNN	0.562	0.537•	0.562	0.437•	0.437•	0.562
	SVM	0.562	0.562	0.473•	0.562	0.437•	0.437•
	GBRT	0.758	0.513•	0.652•	0.566•	0.660•	0.590•
Statistic	Average	0.734	0.663	0.635	0.593	0.634	0.612
	Win/Loss/Tie	108/17/40*	19/5/9	23/3/7	21/2/10	23/4/6	22/3/8
	F-rank	2.10	3.36	3.57	4.42	3.34	4.18

* The total Win/Loss/Tie cases of OSFS-Vague. • The cases than OSFS-Vague wins the other models in comparison. ◦ The cases that OSFS-Vague loses the comparison.

SAOLA and α-investing is significantly higher than that obtained by other algorithms, obviously, the number of features obtained by different datasets was also significantly different. For example, SAOLA obtained 2 features in Colon and 45 features in Lymphoma.

4.2.3 | Prediction accuracy analysis for online feature selection

To prove the advantage of OSFS-Vague compared with other algorithms, we use different classifiers to obtain prediction

TABLE 7 The number of selected features.

Dataset	OSFS-Vague	OFS-Density	OSFS	Fast-OSFS	SAOLA	α-investing	NRS-SFSF
Lung cancer	6	7	5	5	23	54	3
TOX_171	8	6	8	8	10	26	25
Lymphoma	6	5	5	5	45	60	2
Gas sensor array under flow modulation	5	6	2	2	3	5	3
Prostate	11	5	4	4	6	10	7
Lung	6	11	6	6	6	36	9
DLBCL	6	7	5	5	14	19	8
Leukemia	6	2	4	7	13	15	6
Parkinson disease classification	4	2	9	9	7	32	20
Colon	6	8	3	3	2	4	6
Arcene	5	11	8	8	27	33	16
Average	6.3	6.4	5.4	5.6	14.3	26.7	9.5

accuracy, and utilise the F-rank and Wilcoxon signed-ranks tests in statistical tests to verify the credibility of the data in this paper.

OSFS-Vague and other online streaming feature algorithms in hyperparameters of KNN, SVM, and GBRT classifiers are all the same and all algorithms on the 5-fold cross-validation prediction accuracy. Although some results of OSFS-Vague are lower than other algorithms, most results are better than other algorithms in the next to last of Table 8. In particular, the win-loss-tie ratio of OSFS-Vague is 23/2/8, which is the greatest advantage compared with the SAOLA. On the contrary, compared with the OFS-Density, OSFS-Vague has little advantage, which the win-loss-tie ratio is 21/7/5. The F-rank is a widely used statistical technique for contrasting various algorithms. The lower the F-rank value, the greater the advantage of the algorithm. In Table 8, the F-rank value of OSFS-Vague is the smallest, so the method presented in this paper has the highest overall prediction accuracy of the three classifiers.

In addition, to prove whether OSFS-Vague is significantly different compared to other algorithms. We use the Wilcoxon Signed-Ranks to compare OSFS-Vague with other methods, and the outcomes are displayed in Table 9. It mainly includes three results: $R+$, R−, and p-Value. The p-Value is the significance level, which represents the difference between test algorithms. Compared with other algorithms, and the p-Value of OSFS-Vague is all less than 0.05, indicating OSFS-Vague has greater accuracy than its peers.

4.2.4 | Time consumption analysis

Data is complex and dynamic in real life, we selected datasets with many samples or many features in the experiment. It is mainly to observe whether there are differences in time consumption under different data distributions. The specific results are shown in Table 10, more specifically, Lung cancer is the longest dataset consumed by OSFS-Vague and OFS-Density. However, OSFS-Vague consumes less time on the Lung dataset than the DLBCL dataset while OFS-Density consumes more time on the Lung than the DLBCL.

According to the F-rank of the penultimate row in Table 10, SAOLA has the fastest processing speed. It is followed by OSFS-Vague, whose processing time of datasets is less than 1 minute. The processing time of α-investing and NRS-SFSF was also less than 1 minute but longer than the OSFS-Vague. Finally, OFS-Density, OSFS, and Fast-OSFS take the longest time to process datasets.

4.3 | The optimal parameter and weakly redundant features analysis of OSFS-Vague (RQ. 2)

In this part, we will examine how parameter λ affects OSFS-Vague. The values are 0, 0.01, 0.05 and 0.1, respectively. The influence of parameters on the OSFS-Vague is explained in classification accuracy and running duration. In the case of $\lambda = 0.05$, the experiment also tested the changes in algorithm accuracy and training time when Algorithm 3 was removed. To verify the necessity of preserving weakly redundant features.

From Figure 3 and Table 11, the influence of λ on the prediction accuracy of different classifiers. Specifically, SVM has little difference in prediction accuracy of OSFS-Vague with different λ. KNN and GBRT have the highest prediction accuracy when the $\lambda = 0.05$, but it is not obvious. The different hyperparameters of OSFS-Vague have no significant difference in time consumed by feature selection, but the larger the hyperparameters, the more time consumed.

If the parameter $\lambda = 0.1$ is selected, the algorithm will retain the most features. Which may retain more redundant features. This result leads to low accuracy, and the time consumption will increase. If the parameter $\lambda = 0$, the selection feature is too strict, which also influences accuracy. When

TABLE 8 Prediction accuracy of OSFS-Vague vs online streaming feature selection algorithms.

Dataset	Type	OSFS-Vague	OFS-Density	OSFS	Fast-OSFS	SAOLA	α-investing	NRS-SFSF
Lung cancer	KNN	0.839	0.827•	0.827•	0.827•	0.827•	0.827•	0.827•
	SVM	0.172	0.827○	0.172	0.172	0.318○	0.827○	0.827○
	GBRT	0.986	0.827•	0.973•	0.976•	0.979•	0.965•	0.816•
TOX_171	KNN	1	1	1	1	1	1	1
	SVM	0.641	0.729○	0.617•	0.617•	0.488•	0.635•	0.610•
	GBRT	0.555	0.510•	0.426•	0.433•	0.404•	0.423•	0.339•
Lymphoma	KNN	0.672	0.672	0.672	0.672	0.147•	0.212•	0.672
	SVM	0.672	0.672	0.672	0.672	0.147•	0.356•	0.672
	GBRT	0.984	0.934•	0.889•	0.897•	0.865•	0.918•	0.783•
Gas sensor array under flow modulation	KNN	0.991	0.995○	0.8•	0.8•	0.747•	0.921•	0.952•
	SVM	0.776	0.978○	0.344•	0.344•	0.344•	0.874○	0.719•
	GBRT	0.845	0.715•	0.561•	0.549•	0.607•	0.715•	0.650•
Prostate	KNN	0.901	0.485•	0.485•	0.485•	0.514•	0.681•	0.475•
	SVM	0.514	0.485•	0.485•	0.485•	0.514	0.970○	0.514
	GBRT	0.901	0.811•	0.908○	0.908○	0.880•	0.888•	0.579•
Lung	KNN	0.683	0.103•	0.099•	0.099•	0.099•	0.095•	0.683
	SVM	0.683	0.703○	0.103•	0.103•	0.099•	0.683	0.683
	GBRT	0.506	0.824○	0.439•	0.464•	0.535○	0.311•	0.723○
DLBCL	KNN	0.749	0.749	0.749	0.250•	0.250•	0.749	0.250•
	SVM	0.750	0.250•	0.750	0.750	0.750	0.750	0.750
	GBRT	0.888	0.862•	0.897○	0.901○	0.888	0.631•	0.730•
Leukemia	KNN	0.951	0.912•	0.867•	0.968○	0.867•	0.863•	0.889•
	SVM	0.901	0.775•	0.828•	0.929○	0.842•	0.831•	0.700•
	GBRT	0.943	0.897•	0.926•	0.933•	0.922•	0.901•	0.796•
Parkinson disease classification	KNN	0.924	0.791•	0.850•	0.850•	0.850•	0.723•	0.922•
	SVM	0.746	0.746	0.746	0.746	0.746	0.746	0.746
	GBRT	0.748	0.746•	0.746•	0.746•	0.746•	0.746•	0.748
Colon	KNN	1	0.992•	0.996•	0.980•	0.947•	0.992•	0.942•
	SVM	0.914	0.902•	0.885•	0.885•	0.639•	0.918○	0.688•
	GBRT	0.803	0.775•	0.802•	0.786•	0.672•	0.720•	0.548•
Arcene	KNN	0.562	0.473•	0.562	0.562	0.562	0.562	0.562
	SVM	0.562	0.473•	0.473•	0.437•	0.562	0.437•	0.562
	GBRT	0.789	0.790○	0.714•	0.715•	0.789	0.799○	0.694•
Statistic	Average	0.774	0.732	0.674	0.664	0.628	0.717	0.698
	Win/Loss/Tie	131/22/45*	21/7/5	23/2/8	22/4/7	23/2/8	22/5/6	20/2/11
	F-rank	2.35	3.86	4.32	4.14	4.82	3.98	4.53

* The total Win/Loss/Tie cases of OSFS-Vague. • The cases than OSFS-Vague wins the other models in comparison. ○ The cases that OSFS-Vague loses the comparison.

Algorithm 3 is removed, the prediction accuracy under the KNN classifier is unchanged, but the prediction accuracy under the SVM and GBRT classifier decreases. Because of the removal of the replacement feature method, the training speed has increased, but not significantly.

Therefore, a suitable parameter is to achieve a balance between prediction accuracy and time consumption. From Table 11, when parameter $\lambda = 0.05$, prediction accuracy achieves the highest, and time consumption is not significantly improved.

4.4 | Summary of experiments

In this section, the advantages and disadvantages of some algorithms are comprehensively explained through the test of accuracy, number of feature selections, and time consumption.

OFS-Density exhibits the second highest overall prediction accuracy, while it is relatively time-consuming when dealing with the Lung Cancer, TOX_171, and Parkinson disease Classification datasets. As the data volume increases, the processing time of OFS-Density also increases accordingly. In the number of feature selections, there is no discernible difference between OSFS-Vague and OFS-Density.

OSFS is a classical algorithm. The prediction accuracy of OSFS is better than OSFS-Vagure in three results, OSFS is slow in processing large scale data and has no advantage in precision. Moreover, its number of feature selections is more than OSFS-Vague.

TABLE 9 Wilcoxon signed-rank test results.

Comparison	$R+$	R-	p-Value
OSFS-Vague VS ReliefF	226	74	0.0154
OSFS-Vague VS Fisher	320.5	57.5	0.0008
OSFS-Vague VS MI	255	45	0.0014
OSFS-Vague VS PCC	314	64	0.0013
OSFS-Vague VS laplacian	274	51	0.0014
OSFS-Vague VS OFS-density	305	101	0.0104
OSFS-Vague VS OSFS	316	9	0.0001
OSFS-Vague VS Fast-OSFS	324.5	26.5	0.0001
OSFS-Vague VS SAOLA	306	19	0.0001
OSFS-Vague VS α-investing	305	73	0.0027
OSFS-Vague VS NRS-SFSF	215	38	0.0021

Compared with the OSFS, Fast-OSFS optimises the screening process, thus it has faster training time and better prediction accuracy. Fast-OSFS has a higher result accuracy than OSFS. However, it still falls short of OSFS-Vague.

The difference between α-investing and other online streaming feature selections is that hyperparameters are required to be set in advance. Appropriate hyperparameters is able to increase the training speed and the prediction accuracy. However, it is not practical to search the appropriate hyperparameters for each dataset. Compared with OSFS-Vague, α-investing only has advantages in training speed, and its prediction accuracy is not significantly different from OSF-Density.

OSFS-Vague has a good performance in processing both small and large scale data. The number of feature selections is moderate, and there are not too many redundant features to affect the prediction accuracy. We may infer the following conclusions from the outcomes shown above.

- OSFS-Vague has the greatest prediction accuracy on three classifiers. Data results are more reliable by the F-rank and Wilcoxon Signed-Ranks test. At the same time, not all the results of the OSFS-Vague are better than other algorithms. In some datasets, the results of this paper are worse than other algorithms.
- The test of training speed mainly investigates the training speeds for different algorithms in different datasets. Where, the training speed of OFS-Density, OSFS, and Fast-OSFS is relatively slow on large scale data, while OSFS-Vague, SAOLA, and α-investinghold fast processing speed on.
- Too few or too many features will affect the prediction accuracy. In the experiment, α-investing requires to preset hyperparameters in advance for training. Thus, the

TABLE 10 Time consumption comparison.

Dataset	OSFS-Vague	OFS-Density	OSFS	Fast-OSFS	SAOLA	α-investing	NRS-SFSF
Lung cancer	25.32	215.91•	3871.15•	5685.01•	9.54◦	37.02•	16.68◦
TOX_171	16.10	102.86•	1191.78•	1350.21•	2.53◦	26.57•	7.63◦
Lymphoma	3.73	10.46•	531.76•	523.35•	5.97•	11.31•	1.25◦
Gas sensor array under flow modulation	0.727	1.60•	1.14•	1.11•	0.30◦	0.41◦	0.89•
Prostate	6.63	34.93•	85.97•	83.42•	2.93◦	5.68◦	2.96◦
Lung	8.48	82.26•	907.01•	916.87•	1.59◦	29.30•	5.49◦
DLBCL	9.73	19.03•	217.02•	216.38•	2.07◦	6.37◦	2.81◦
Leukemia	3.18	13.81•	590.36•	615.15•	6.15•	11.29•	3.16◦
Parkinson disease classification	5.01	400.49•	1440.43•	864.51•	0.38◦	5.12•	7.76•
Colon	1.14	6.55•	6.21•	6.18•	0.66◦	2.14•	1.02◦
Arcene	17.99	332.62•	3410.04•	3399.09•	4.29◦	14.27◦	14.77◦
Win/loss	44/22	11/0	11/0	11/0	2/9	7/4	2/9
F-rank	3	5.27	6.45	6.18	1.36	3.45	2.27

• The cases than OSFS-Vague wins the other models in comparison. ◦ The cases that OSFS-Vague loses the comparison.

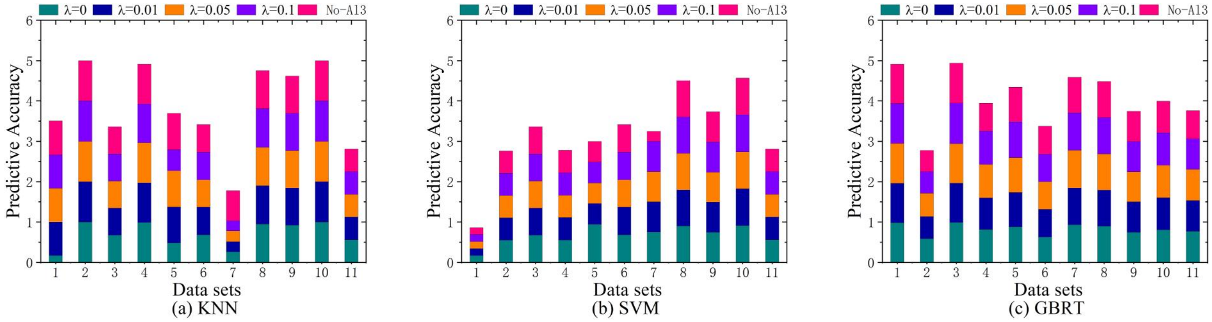

FIGURE 3 Predictive accuracy in classifiers varyes with four different values of λ.

TABLE 11 The mean values of different λ on predictive accuracy and running time.

	$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.1$	No-AL3
KNN classifier	0.744	0.840	0.843	0.758	0.843
SVM classifier	0.677	0.638	0.638	0.638	0.593
GBRT classifier	0.820	0.815	0.824	0.821	0.795
Running time	9.927	9.837	9.948	10.307	9.588

number of features is not controllable, and the quantity of feature selection acquired by different methods does not differ much.

To sum up, OSFS-Vague is superior to other algorithms under the comprehensive results. Firstly, most prediction accuracy results are better than other algorithms. Secondly, results can be obtained quickly regardless of data size. Finally, the quantity of feature selection is moderate, and the accuracy will not be reduced due to too many or too few features.

However, the OSFS-Vague method has certain limitations. Firstly, OSFS-Vague is unable to make an effective filter when samples are missing in some features of the dataset. Secondly, the OSFS-Vague method only considers the relationship between a single feature and labels when screening features, but does not consider the relationship between multiple features and labels. Because all the features are calculated in this way, the training time will be unacceptable. Thirdly, when the number of features in the dataset is large, the number of weakly redundant features obtained by OSFS-Vague will also increase, resulting in a longer training time for subsequent filters. In summary, OSFS-Vague still has many places that require improvement, which is our future research direction.

5 | CONCLUSION

OSFS-Vague is able to capture the relationship between features and labels by analyzing their distribution. OSFS-Vague eliminates the need to pre-set any hyperparameters and offers a better understanding of the importance of features in the datasets. Additionally, a screening process is introduced for weakly redundant features that follow the principle of three-way decision, which is consistent with human cognitive processes. In the experiment, the quantity of features obtained by the OSFS-Vague is moderate, then the training time is better than most algorithms, and the prediction accuracy is higher than all other methods.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China (Grant number 62066049, Grant number 62221005, Grant number 61936001, Grant number 62176070), the Guizhou Provincial Department of Education Colleges and Universities Science and Technology Innovation Team (QJJ [2023]084), Excellent Young Scientific and Technological Talents Foundation of Guizhou Province (QKH-platform talent (2021) 5627), Science and Technology Top Talent Project of Guizhou Education Department (QJJ2022[088]), Science and Technology Project of Zunyi (ZSKRPT[2023] 3), in part by the Science and Technology Foundation of State Grid Corporation of China under grant 1400-202357341A-1-1-ZN, in part by the Chongqing Technical Innovation and Application Development Special Project under grant CSTB2023TIAD-KPX0037.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are included within the article.

ORCID

Jie Yang https://orcid.org/0000-0002-6580-9287
Zhijun Wang https://orcid.org/0009-0007-0447-3795

REFERENCES

1. He, Y., et al.: Toward mining capricious data streams: a generative approach. IEEE Transact. Neural Networks Learn. Syst. 32(3), 1228–1240 (2020). https://doi.org/10.1109/tnnls.2020.2981386
2. He, Y., et al.: Online learning in variable feature spaces under incomplete supervision. Proc. AAAI Conf. Artif. Intell. 35(5), 4106–4114 (2021). https://doi.org/10.1609/aaai.v35i5.16532
3. Wu, D., et al.: A latent factor analysis-based approach to online sparse streaming feature selection. IEEE Trans. Syst. Man. Cybern. 52(11), 6744–6758 (2022). https://doi.org/10.1109/tsmc.2021.3096065
4. Yu, K., Wu, X.D., Pei, J.: Scalable and accurate online feature selection for big data. ACM Trans. Knowl. Discov. Data 11(2), 1–39 (2016). https://doi.org/10.1145/2976744

5. Perkins, S., Theiler, J.: Online feature selection using grafting. In: Proceedings of the Twentieth International Conference on International Conference on Machine Learning, pp. 592–599 (2003)
6. Wu, X.D., et al.: Online feature selection with streaming features. IEEE Trans. Pattern Anal. Mach. Intell. 35(5), 1178–1192 (2013). https://doi.org/10.1109/tpami.2012.197
7. Zhou, P., Wang, N., Zhao, S.: Online group streaming feature selection considering feature interaction. Knowl. Base Syst. 226, 1–11 (2021). https://doi.org/10.1016/j.knosys.2021.107157
8. Nuaimi, N.A., Masud, M.M.: Online streaming feature selection with incremental feature grouping. Data Min. Knowl. Discov. 10(4), 1–23 (2019)
9. Zhou, P., et al.: Online feature selection for high-dimensional class-imbalanced data. Knowl. Base Syst. 136(15), 187–199 (2017). https://doi.org/10.1016/j.knosys.2017.09.006
10. Zhou, J., et al.: Streamwise feature selection. J. Mach. Learn. Res. 7(9), 1861–1885 (2006)
11. Li, J.D., et al.: Feature selection: a data perspective. ACM Comput. Surv. 50(6), 1–45 (2017). https://doi.org/10.1145/3136625
12. Ou, X.J., et al.: Online causal feature selection for streaming features. IEEE Transact. Neural Networks Learn. Syst. 34(3), 1563–1577 (2021)
13. You, D.L., et al.: Local causal structure learning for streaming features. Inf. Sci. 647, 119502 (2023). https://doi.org/10.1016/j.ins.2023.119502
14. Wu, D., et al.: Online semi-supervised learning with mix-typed streaming features. Proc. AAAI Conf. Artif. Intell. 37(4), 4720–4728 (2023). https://doi.org/10.1609/aaai.v37i4.25596
15. Zhou, P., Zhang, Y.Y., Wu, X.: General assembly framework for online streaming feature selection via Rough Set models. Expert Syst. Appl. 204, 1–14 (2022). https://doi.org/10.1016/j.eswa.2022.117520
16. Zhang, X., et al.: Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy. Pattern Recogn. 56, 1–15 (2016). https://doi.org/10.1016/j.patcog.2016.02.013
17. Liu, J.H., et al.: Online multi-label streaming feature selection based on neighborhood rough set. Pattern Recogn. 84, 273–287 (2018). https://doi.org/10.1016/j.patcog.2018.07.021
18. Li, S.J., et al.: Online streaming feature selection based on neighborhood rough set. Appl. Soft Comput. 113, 1–19 (2021). https://doi.org/10.1016/j.asoc.2021.108025
19. Vluymans, S., et al.: Dynamic affinity-based classification of multi-class imbalanced data with one-versus-one decomposition: a fuzzy rough set approach. Knowl. Inf. Syst. 56(1), 55–84 (2018). https://doi.org/10.1007/s10115-017-1126-1
20. Eskandari, S., Javidi, M.M.: Online streaming feature selection using rough sets. Int. J. Approx. Reason. 69, 35–57 (2016). https://doi.org/10.1016/j.ijar.2015.11.006
21. Zhou, P., et al.: OFS-Density: a novel online streaming feature selection method. Pattern Recogn. 86(1), 48–61 (2019). https://doi.org/10.1016/j.patcog.2018.08.009
22. Zhou, P., et al.: Online early terminated streaming feature selection based on Rough Set theory. Appl. Soft Comput. 113, 1–12 (2021). https://doi.org/10.1016/j.asoc.2021.107993
23. Wang, Z.Q., et al.: Information-Theory-based nondominated sorting ant colony optimization for multiobjective feature selection in classification. IEEE Trans. Cybern. 53(8), 5276–5289 (2023). https://doi.org/10.1109/tcyb.2022.3185554
24. Gong, Y.L., et al.: A length-adaptive non-dominated sorting genetic algorithm for Bi-objective HighDimensional feature selection. IEEE/CAA J. Autom. Sin. 10(9), 1834–1844 (2023). https://doi.org/10.1109/jas.2023.123648
25. Han, S.F., et al.: Locating multiple equivalent feature subsets in feature selection for imbalanced classification. IEEE Trans. Knowl. Data Eng. 35(9), 9195–9209 (2023). https://doi.org/10.1109/tkde.2022.3222047
26. Han, S.F., et al.: Information-utilization-method-assisted multimodal multiobjective optimization and application to credit card fraud detection. IEEE Trans. Comput. Soc. Syst. 8(4), 856–869 (2021). https://doi.org/10.1109/tcss.2021.3061439
27. Han, S.F., et al.: Competition-driven multimodal multiobjective optimization and its application to feature selection for credit card fraud detection. IEEE Trans. Syst. Man. Cybern. 52(12), 7845–7857 (2012). https://doi.org/10.1109/tsmc.2022.3171549
28. Gau, W.L., Buehrer, D.J.: Vague sets. IEEE Trans. Syst. 23(2), 610–614 (1993). https://doi.org/10.1109/21.229476
29. Yao, Y.Y.: Three-way decisions with probabilistic rough sets. Inf. Sci. 180(3), 341–353 (2010). https://doi.org/10.1016/j.ins.2009.09.021
30. Yao, Y.Y.: The geometry of three-way decision. Appl. Intell. 51(9), 6298–6325 (2021). https://doi.org/10.1007/s10489-020-02142-z
31. Li, T., et al.: Recent advances in feature selection and its applications. Knowl. Inf. Syst. 53(3), 551–577 (2017). https://doi.org/10.1007/s10115-017-1059-8
32. Cai, J., et al.: Feature selection in machine learning: a new perspective. Neurocomputing 300(26), 70–79 (2018). https://doi.org/10.1016/j.neucom.2017.11.077
33. You, D., et al.: Online learning from incomplete and imbalanced data streams. IEEE Trans. Knowl. Data Eng. 35(10), 10650–10665 (2003). https://doi.org/10.1109/tkde.2023.3250472
34. Zhou, P., et al.: A new online feature selection method using neighborhood rough set. In: IEEE International Conference on Big Knowledge, pp. 135–142 (2017)
35. Sikonja, R.M.M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. 53(1-2), 23–69 (2003). https://doi.org/10.1023/a:1025667309714
36. He, X., et al.: Laplacian score for feature selection. Adv. Neural Inform. 17, 507–514 (2005)
37. Gu, Q., et al.: Generalized Fisher score for feature selection. In: Proceedings of Conference on Uai (2011)
38. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. 24(1), 175–186 (2014). https://doi.org/10.1007/s00521-013-1368-0
39. Roffo, G., et al.: Infinite feature selection. Proc. IEEE Int. Conf. Comput., 4202–4210 (2015)
40. Tsai, C.F., Sung, Y.T.: Ensemble feature selection in high dimension, low sample size datasets: parallel and serial combination approaches. Knowl. Base Syst. 203(5), 106097 (2020). https://doi.org/10.1016/j.knosys.2020.106097
41. Nadimi-Shahraki, M.H., Zamani, H., Mirjalili, S.: Enhanced whale optimization algorithm for medical feature selection: a COVID-19 case study. Comput. Biol. Med. 148, 105858 (2022). https://doi.org/10.1016/j.compbiomed.2022.105858
42. Hu, R.Y., et al.: Multi-task multi-modality SVM for early COVID-19 Diagnosis using chest CT data. Inf. Process. Manag. 59(1), 102782 (2022). https://doi.org/10.1016/j.ipm.2021.102782
43. Yang, Y.Y., et al.: Incremental feature selection by sample selection and feature-based accelerator. Appl. Soft Comput. 121, 108800 (2022). https://doi.org/10.1016/j.asoc.2022.108800
44. Zhou, P., et al.: Balanced spectral feature selection. IEEE Trans. Cybern. 53(7), 4232–4244 (2023). https://doi.org/10.1109/tcyb.2022.3160244
45. Barddal, J.P., et al.: Boosting decision stumps for dynamic feature selection on data streams. Inf. Syst. 83, 13–29 (2019). https://doi.org/10.1016/j.is.2019.02.003
46. de Moraes, M.B., Sapaio, A.L.S.: A comparative study of feature selection methods for binary text streams classification. Evolving Systems 12(4), 997–1013 (2021). https://doi.org/10.1007/s12530-020-09357-y
47. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
48. Zadeh, L.A.: Fuzzy sets. Information and control. Inf. Control 8(3), 338–353 (1965). https://doi.org/10.1016/s0019-9958(65)90241-x
49. Székely, G.J., Rizzo, M.F., Bakirov, N.K.: Measuring and testing dependence by correlation of distance. Ann. Stat. 35(6), 2769–2794 (2008). https://doi.org/10.1214/009053607000000505
50. Xu, H.H., Deng, Y.: Dependent evidence combination based on shearman coefficient and Pearson coefficient. IEEE Access 6, 11634–11640 (2018). https://doi.org/10.1109/access.2017.2783320
51. Yu, K., et al.: Classification with streaming features: an emerging-pattern mining approach. ACM Trans. Knowl. Discov. Data 9(4), 1–31 (2015). https://doi.org/10.1145/2700409
52. Ling, Z.L., et al.: BAMB: a balanced markov blanket discovery approach to feature selection. ACM Trans. Intell. Syst. Technol. 10(5), 1–25 (2019). https://doi.org/10.1145/3335676

53. Wang, H., et al.: Towards efficient and effective discovery of Markov blankets for feature selection. Inf. Sci. 509, 227–242 (2020). https://doi.org/10.1016/j.ins.2019.09.010
54. Zhang, Q.H., et al.: Democratic three-way decisions based on voting mechanism. Int. J. Mach. Learn. Cybern 13(1), 99–114 (2021). https://doi.org/10.1007/s13042-021-01367-9
55. Asuncion, A., Newman, D.: UCI Machine Learning Repository (2007). http://archive.ics.uci.edu/ml
56. Yang, K., et al.: A stable gene selection in microarray data analysis. BMC Bioinf. 7(28), 1–16 (2006). https://doi.org/10.1186/1471-2105-7-228
57. Wasikowski, M., Chen, X.: Combating the small sample class imbalance problem using feature selection. IEEE Trans. Knowl. Data Eng. 22(10), 1388–1400 (2010). https://doi.org/10.1109/tkde.2009.187
58. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)

How to cite this article: Yang, J., et al.: OSFS-Vague: Online streaming feature selection algorithm based on vague set. CAAI Trans. Intell. Technol. 9(6), 1451–1466 (2024). https://doi.org/10.1049/cit2.12327

APPENDIX

The OSFS-Vague in this paper is available from the following website: https://github.com/pnkx/Online-feature-streaming-selection-method.git.